npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Multi-component background learning automates signal detection for spectroscopic data

Sebastian E. Ament[1], Helge S. Stein[2], Dan Guevarra[2], Lan Zhou[2], Joel A. Haber[2], David A. Boyd[2], Mitsutaro Umehara[2,3], John M. Gregoire[2] and Carla P. Gomes[1]

Automated experimentation has yielded data acquisition rates that supersede human processing capabilities. Artificial Intelligence offers new possibilities for automating data interpretation to generate large, high-quality datasets. Background subtraction is a long-standing challenge, particularly in settings where multiple sources of the background signal coexist, and automatic extraction of signals of interest from measured signals accelerates data interpretation. Herein, we present an unsupervised probabilistic learning approach that analyzes large data collections to identify multiple background sources and establish the probability that any given data point contains a signal of interest. The approach is demonstrated on X-ray diffraction and Raman spectroscopy data and is suitable to any type of data where the signal of interest is a positive addition to the background signals. While the model can incorporate prior knowledge, it does not require knowledge of the signals since the shapes of the background signals, the noise levels, and the signal of interest are simultaneously learned via a probabilistic matrix factorization framework. Automated identification of interpretable signals by unsupervised probabilistic learning avoids the injection of human bias and expedites signal extraction in large datasets, a transformative capability with many applications in the physical sciences and beyond.

npj Computational Materials (2019)5:77 ; https://doi.org/10.1038/s41524-019-0213-0

INTRODUCTION

Data analysis and interpretation are pervasive in physical sciences research and typically involve information extraction from noisy and background-containing signals.[1–3] Examples from materials science include the identification of crystal structures from X-ray diffraction patterns[4] and chemical species from X-ray photoelectron spectra.[5] Distinguishing the signal of interest from background signals comprises a major hurdle, and any errors in making these distinctions can alter data interpretation.[6,7] The identification of the signal of interest often requires expert knowledge[8,9] and/or application of empirical algorithms, motivating the establishment of a more principled approach.

An example of principled background removal in physical sciences concerns the Bremsstrahlung radiation observed in energy-dispersive X-ray spectroscopy (EDS),[10,11] which provides an ideal situation for background identification because there is a single primary background source whose shape can be derived from fundamental physics.[10–13] On the other hand, measurements such as X-ray diffraction (XRD) typically involve a variety of background sources. The background sources of measured X-ray intensities can include scattering by air, elastic scattering by the sample, and scattering by the substrate or sample support, which appear in the detector signal in combination with the desired inelastic scattering from the sample of interest. Furthermore, a given background signal may be attenuated differently over a set of measurements, but it always provides a non-zero contribution to the measured signal. Since the level of these different background signals can vary independently, it is not possible to identify a single characteristic background pattern, motivating the establishment of a multi-component model. Raman spectroscopy similarly involves a variety of background sources. Herein, XRD and Raman data are used as specific examples in which the measured signal is the combination of positive intensities including the signal of interest and any number of background signals.

Empirical background subtraction models[6,7,14,15] typically require manual fine tuning of parameters. For example, the XRD background subtraction algorithm from Sonneveld and Visser[6] requires parameters for the smoothness of the data and the magnitude of the intensity gradients for peaks of interest. Though the algorithm can be implemented effectively, as reflected by its incorporation into several commercial software packages for XRD analysis, users still need to fine-tune the parameters to avoid distortion of the peaks of interest and overestimation of the background signal.

Further, as is shown in the current work, there are complex background signals which defy approaches based on fitting a background model to a single spectrogram at a time. More recently, background identification through analysis of a collection of measurements has been performed using methods such as principal component analysis (PCA)[16] or polynomial fitting,[15] which still require expert knowledge in discriminating background from signal and do not guarantee non-negativity of the extracted signal.

We introduce Multi-Component Background Learning (MCBL), a fundamentally new approach to background subtraction and signal identification. MCBL leverages the power of big data by inferring background and signals of interest from an entire dataset of spectrograms. Second, MCBL's inference task is enabled by a

[1]Department of Computer Science, Cornell University, Ithaca, NY 14850, USA; [2]Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, CA 91125, USA and [3]Future Mobility Research Department, Toyota Research Institute of North America, Ann Arbor, MI 48105, USA
Correspondence: John M. Gregoire (gregoire@caltech.edu) or Carla P. Gomes (gomes@cs.cornell.edu)

Received: 18 February 2019 Accepted: 26 June 2019
Published online: 19 July 2019

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

novel probabilistic generative model of the spectroscopic data where the background components, the noise variance, and the level of spectroscopic activity are all concomitantly learned from the data. The comprehensiveness of the learning model is key for achieving autonomous interpretation of spectroscopic data, a goal of increasing practical importance for emerging technologies such as materials acceleration platforms.[3] Third, MCBL provides the probability that any given data point contains a (non-background) signal of interest. This probability is automatically inferred by the algorithm based on its unified probabilistic framework, and does not rely on human parameter estimates.

Furthermore, the MCBL model is flexible enough to incorporate prior knowledge of different types of background sources.

For example, a common assumption is the smoothness of the background signals, which the algorithm can incorporate by enforcing a user-defined smoothness constraint. Note however, that the algorithm is less sensitive to these types of human inputs than other algorithms, especially when the algorithm is given a large number of spectrograms. Providing prior knowledge is especially important in challenging cases where there are many complex background signals and data are scarce. Last, the MCBL algorithm requires a noise model. We describe its principled design for XRD and Raman data, as well as the physical meaning of each parameter, in the Methods section. In addition, the noise model's parameters are not required to be chosen manually but can be learned from the data.

MCBL is demonstrated using large datasets from two common techniques in materials characterization: XRD and Raman spectroscopy. In both cases, the data were acquired using composition libraries that were synthesized to measure and identify composition-structure-property relationships,[17] a central tenet of combinatorial materials science.[18] Automated inference of the crystal structures from XRD or Raman characterization of the composition library, i.e. "Phase Mapping", is a long-standing bottleneck in materials discovery.[8] Phase Mapping algorithms have been plagued by both insufficient background removal and incorrect labeling of signals of interest as background or noise. Unsupervised, principled background removal circumvents these issues to increase both the speed and the quality of data interpretation.

RESULTS

X-ray diffraction

To demonstrate the performance of MCBL and illustrate some of the more subtle aspects of the model and its deployment, we apply it to a particularly challenging XRD example in which there are multiple background sources, including a background source whose intensity is substantially higher than the signal of interest. In this case, the strong background signal is from diffraction of the SnO_2 in the substrate, introducing unwanted peaks into the dataset that are quite similar in shape to those in the desired signal from the thin film sample. Furthermore, over a series of 186 reflection-geometry measurements on different thin film compositions, the variable density and thickness of the thin film of interest alters the shape and intensity of the substrate signal. Provided that the set of 186 samples contains more variability in the signal of interest than the background signal, which it does due to the variety of crystal structures in the 186 unique compositions, MCBL identifies the unique combination of background signals for each of the 186 measured diffraction patterns. Note that we have prior knowledge that there are two distinct types of background sources: diffraction signals from the crystalline substrate and smoothly varying signals from other sources including elastic scattering and air scattering. We inject this knowledge into the model by allowing one type of background component to have intensity only in the vicinity of known substrate diffraction peaks (scattering vector magnitudes 18.5–19.2, 23.6–24.1, and 26.2–26.8 nm^{-1}), while the other type of background component is enforced to be smoothly varying.

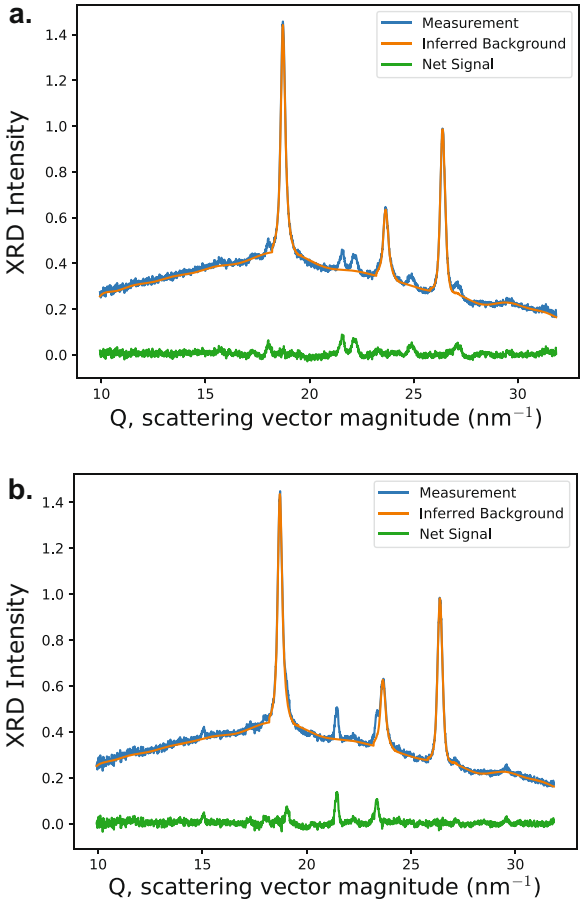

Fig. 1 Two representative samples from the 186 XRD measurements that were collectively used to establish the multi-component background. Each measured signal is shown along with the inferred background, which contains 3 diffraction peaks from the substrate that are much larger in intensity than those of the sample of interest. This measurement-specific inferred background produces a net signal that intentionally retains the measurement noise so that the signal from each sample of interest can be interpreted in the context of the measurement noise. In **a.** a series of relatively small peaks are recovered from the background-dominated signal. In **b.** similar signal recovery is obtained even when peaks of interest strongly overlap with peaks in the background signal

As shown in Fig. 1, the MCBL model identifies the background signal, enabling retention of the desired signal even when the Bragg peak from the sample strongly overlaps that of the substrate. The recovery of the desired signal from the shoulder of the much more intense background signal, as exemplified by the peak near 23 nm^{-1} in Fig. 1b, is uniquely enabled by the model's ability to learn the background signal from the collection of measurements. It is also worth noting that the background models in these 2 examples are different in slight but important ways because the total background signal is unique to each measurement, which is illustrated further in the Raman example below.

Raman spectroscopy

Continued demonstration of MCBL proceeds with a Raman spectroscopy dataset where 2121 metal oxide samples spanning 15 pseudo-quaternary metal oxide composition spaces (5

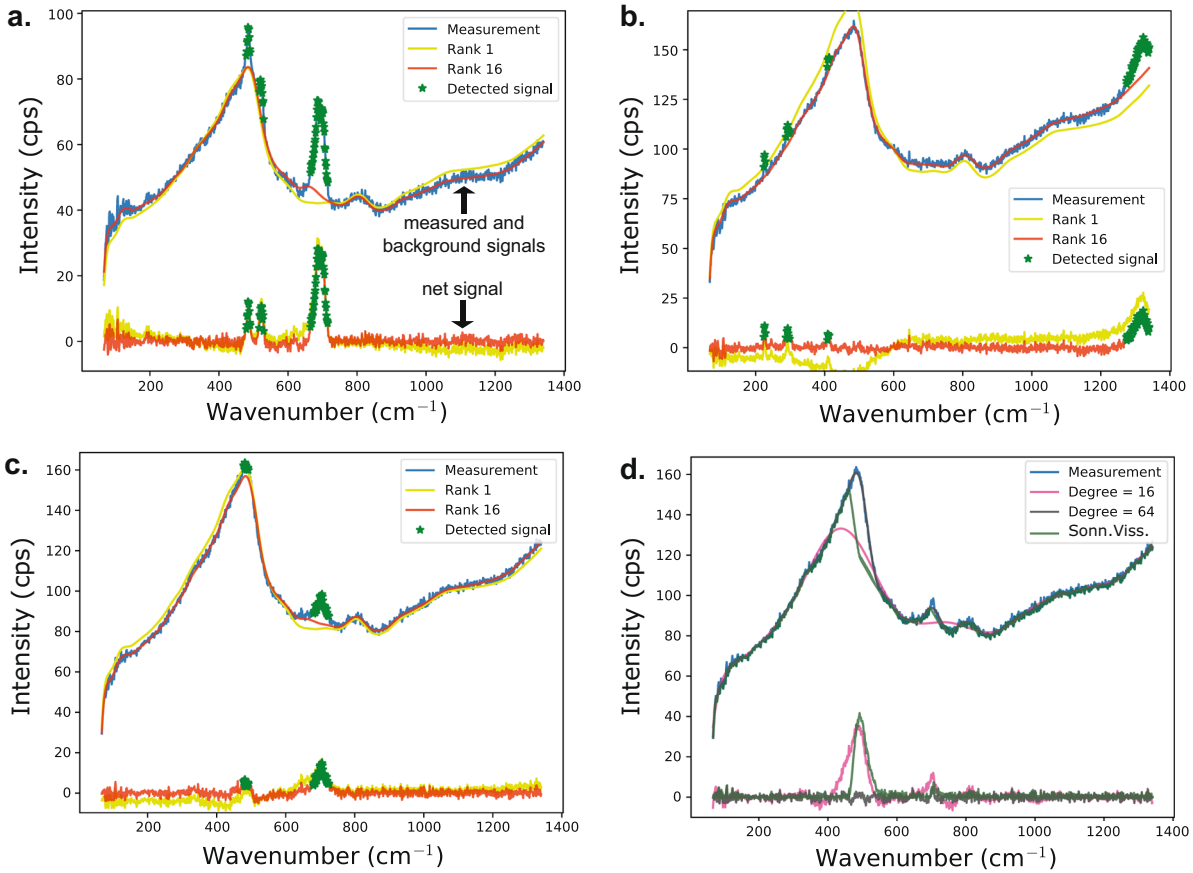

Fig. 2 Three samples (**a**–**c**) from the 2121 Raman measurements representing measurements with weak signals from the sample of interest. Each measured signal is shown along with the inferred background using both 1 background component and 16 background components (rank). The net signals from these 2 background models are also shown at the bottom of each figure, and the data points that are determined by the probablistic model to likely contain signal from the sample are denoted by green stars in both the raw measurement and the net signal from the rank 16 background model. The measurement of **c**. is also shown in **d**. where a traditional polynomial background analysis is performed (see ref. [15] for details) using both 16 and 64-degree polynomials as well the Sonneveld and Visser algorithm.[6] The resulting net signals at the bottom of the figure fail to reflect the signal from the sample

elements including oxygen but systematic variation of the concentrations of only the 4 metals yields dimensionality of a quaternary composition space) were measured using a rapid Raman scanning technique described previously.[19] Similar to the XRD dataset, the Raman signal from the substrate varies in intensity with sample composition, and the high sensitivity of Raman detectors to environmental factors such as room temperature introduces additional variability in background signal. Data acquisition proceeded over a week, during which time-dependent variation in signal levels were observed. These occur, for example, due to day to night temperature variation in the laboratory. While we expect the background to be smooth, a closed mathematical expression is not available, making this dataset well matched to the capabilities of the MCBL model. As discussed in the Methods section, limiting each of the background signals to be smooth makes the results relatively insensitive to the number of background sources included in the model, provided this number is at least as large as the true number of background sources. Since we expect that several sources may be present, 16 is a convenient upper bound and is a standard value to use for datasets where more specific knowledge of the background sources is unavailable.

Since peak shapes, in particular peak widths, are more variable in Raman measurements compared to XRD measurements, and the intensity of the Raman signal of interest is often comparable to the measurement noise, even background—Raman signals are not readily interpretable without additional information. MCBL provides such additional information, in particular the probability that each individual data point contains signal from the sample, i.e., intensity that is not explainable by the background and noise models. For each measured signal, the algorithm produces a probability signal that can be used to reason about the data in subsequent analysis. Since single-point outliers in the measured signals can cause single point outliers in the probablity signal, MCBL factors in the prior knowledge that any Raman feature of interest will span several data points by smoothing the probability signal via kernel regression[20] with a Gaussian kernel of (σ) three data points. Thresholding the smoothed probability signals at 50% provides identification of each data point that likely contains signal from sample of interest.

Representative examples of background identification and removal are shown for three Raman measurements in Fig. 2a–c. Using MCBL with 16 background components yields background-subtracted signals with a flat, near-zero baseline atop which the small signal peaks are far more evident than in the raw data. Since each net signal contains measurement noise, the visual identification of peaks can be assessed in the context of this noise. The results of the probability signal analysis are shown with demarcation of each data point that likely contains signal from the substrate. It is worth noting that researchers often apply smoothing to assist in identification of such small peaks in the signal, although the propensity for modification of the true signal and possibilities for both false positive and false negative peak detection highlights the benefits of the identifying the

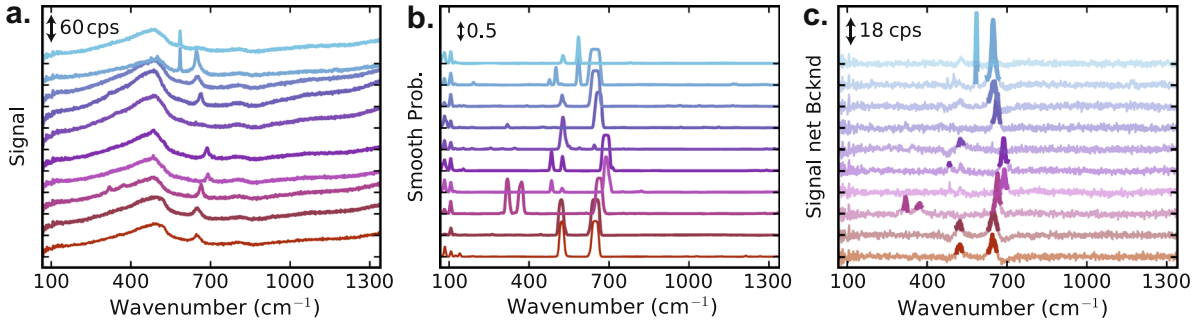

Fig. 3 Stack plots containing 10 measurements from the set of 2121 measured Raman signals, chosen based on variation in observed Raman signal from the sample. The measured signals in **a**. were analyzed collectively with the full dataset to derive a comprehensive background model, including the probability signals indicating the likelihood that each data point contains signal from the sample (**b**), and the background-subtracted signals (**c**). These latter signals contain the modeled signal from the sample, as well as measured noise, and the opaque data points are those whose probability signal is above the 0.5 threshold, providing the user a clear visualization of the signals of interest

background signal using a probabilistic model that considers both the noise and the signal from the sample.

Figure 2a includes examples of peaks from the sample that are notoriously difficult to identify. The peak near 480 cm^{-1} appears atop of a larger peak in the background signal, and the peak near 510 cm^{-1} lies on a strongly sloped portion of the background signal. The intensity at the right edge of the measured signal in Fig. 2a is increasing, so inspecting this individual pattern could not definitively identify that portion of the signal as being absent or inclusive of signal from the sample. The probabilistic model makes this assessment, where no signal from the sample is identified in this portion of Fig. 2a. A sample peak is detected in the analogous portion of the measurement in Fig. 2b where the partial measurement of the peak atop a sloped background would be problematic for any peak fitting (regression)-based search for sample peaks. Figure 2b also demonstrates the importance of the multi-component aspect of the background model. While the background signal is qualitatively similar to the other samples, the quantitative differences that are emblematic of the unique mixture of the background sources render the single-component model unable to provide a clean background-subtracted signal. The model's detection of two peaks in the measured signal of Fig. 2c (near 480 and 680 cm^{-1}) is particularly impressive as even expert manual analysis may hesitate to label these features as sample peaks due to the poor signal-to-noise ratio. Their detection in the probabilistic model is aided by the appearance of the peaks in other measurements, including that of Fig. 2a.

To highlight the quality of the net signals produced by the MCBL model, the measurement of Fig. 2c is shown in Fig. 2d along with traditional polynomial baseline modeling. The lower-order polynomial yields a net signal where the largest peak is actually from a background source, and increasing the polynomial order to capture this feature in the background model results in removal of practically all signal from the sample.

To further illustrate the background removal and peak identification process, Fig. 3 includes a series of ten of the Raman measurements with a variety of peak locations, shapes, and relationships to the background signal. Since the signal probability is calculated for every data point, the probability signals can be plotted in the same manner as the measured signals, as shown in Fig. 3b. The background-subtracted samples in Fig. 3c are shown with partial transparency where the probability signal is below the 50% threshold so that the regions of each pattern that likely contain signal from the sample are highlighted. The sharp, intense peaks in the top two patterns may be easily identified by a variety of algorithms, although identification of many of the broader, weaker features from each sample require the excellent background identification and probabilistic reasoning of the MCBL model.

Probabilistic classification of sample signal and enumeration of background sources

Figure 2b also illustrates a subtle consequence of the model's collective learning of the background signals, measurement noise, and probabilities via the probabilistic framework. The rank 16 model identifies the appropriate background and consequently correctly learns the measurement noise to identify three small peaks between 220 and 420 cm^{-1}. The rank 1 background model is imperfect, and the collection of samples with incorrect background signals inflates the model's estimation of the noise level such that the resulting probability signals do not identify any of these three peaks as likely containing signal from the sample. The comprehensive probabilistic framework enables simultaneous learning of multiple properties of the measured signals, but using a background rank smaller than the true number of background sources is deleterious not only to background removal but also to automated detection of signals of interest.

The classification of measured signals as lacking or containing a signal of interest has a variety of applications ranging from materials discovery to characterization of the background sources. Using the rank 16 background model, 743 of the 2121 measured signals contain at least one datapoint that is likely to contain signal of interest. Using this as the baseline classification of absence or presence of signal from the sample, the performance of lower-rank models can be assessed via the recall (the fraction of the 743 patterns with signal that are correctly identified as having signal) and the precision (the fraction of signals with detected signal that actually have signal). The results are summarized in Fig. 4a and demonstrate the poor performance of the rank 1 model for this classification task, which is due to a confluence of phenomena including that noted above; non-removed background signal can be interpreted as signal of interest (false positive), and the inflated noise level in the noise model can fail to identify small signals of interest (false negative). Increasing to rank 2 greatly improves the recall but not the precision, and increasing to rank 4 largely removes the disparity between recall and precision. Since there is no substantial change upon increasing to rank 8, these results collectively indicate that the number of background sources is three or four. It is worth noting that multiple components are needed to model a single background source if its signal varies in shape over the dataset, so this interpretation of rank as determine the number of sources includes the number of unique physical phenomena that alter the shape of a background signal. The background sources can be further characterized using the wealth

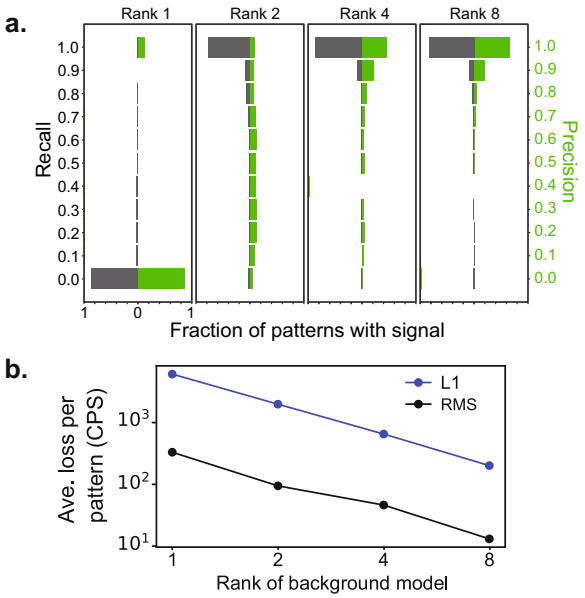

Fig. 4 Background model results for rank 1, 2, 4, and 8 (the number of background sources in the model) are compared to the results from rank 16. In **a**, the probabilistic model is used to classify each measured signal as lacking or having signal from the sample, and the recall and precision of the lower rank models for this classification is shown with a series of histograms all plotted on the same scale. In **b**, each background-subtracted signal is compared to that of the rank 16 model, producing the average L_1 and RMS loss per measured signal

of information provided by the MCBL model, such as the spatial or temporal variation in the intensity of each background source.

Figure 4b includes a similar analysis for how the background-subtracted signals vary with rank. Once again using the rank 16 results as the baseline for comparison, the difference of each background-subtracted signal is measured using both the ℓ_1 and root mean squared (RMS) loss. The average per-signal loss appears to follow a power law relationship with the model rank. Each pattern contains 1023 data points, so starting at rank 4 the ℓ_1 value per data point is about 1 CPS or lower, and comparison to the signals in Fig. 2 demonstrate that this is within the measurement noise, in agreement with the above observation that rank 3 or 4 is sufficient to model the background in this dataset. Using a larger rank has no substantial influence on the resulting signals of interest. This stability in the model's solution is an important feature for unsupervised deployment.

DISCUSSION

The results above demonstrate not only successful background removal but also the generation of insightful probabilistic models for both XRD and Raman data. While background removal is often considered a non-scientific aspect of data interpretation, consider instead the concept that the scientific merit of a chain of analyses is only as strong as its weakest link. Artifacts injected from non-principled background subtraction are inherited by subsequent analyses and can contaminate the scientific interpretation of the data. In general, any modification to measured data should be performed in a manner that reflects a fundamental understanding of the underlying physical processes that give rise to the measured signals. In the present work, this understanding is incorporated with specificity through the establishment of a probability density model for the signal of interest, yet through its parameterization the model retains generality for any measurements involving addition of non-negative sources. A desirable consequence of this principled parameterization of the background model is that the learned parameters provide statistical characterizations of the data, which was demonstrated with analysis of the probability signals and the identification of the number of background sources in the Raman dataset. While not discussed in the present work, after identification of the number of background sources, the individual background signals can be analyzed to study the background sources themselves, and the activations of each of these sources in a dataset enables quantification of the variability in each background source's intensity. While one goal of the algorithm is the generation of background-free signals, these examples illustrate the broader application of the probabilistic learning approach, that the optimized probabilistic model contains deep information about every component of the measured signal.

A principled approach to the identification, removal and statistical evaluation of background signals is established for any measurement type where each measured signal is a combination of non-negative contributions from multiple sources. Through design of a parameterized probability density function for the measured intensities of a signal of interest, a probabilistic framework is established for unsupervised learning of background signals, in particular when there are multiple sources of background whose contributions to the measured signal vary among the set of measurements. In addition to unsupervised operation, the model provides a variety of methods for incorporating prior knowledge, which is demonstrated with an example XRD dataset in which the crystalline substrate produces more intense diffraction patterns than the sample of interest. The probability signals, which indicate where the signal of interest is likely present, are demonstrated using a Raman dataset in which the ~4 background sources are identified and modeled for each measurement, providing signals for further analysis that contain negligible contributions from the background. The probability signals and other parameters can be employed by subsequent reasoning and learning algorithms, making the algorithm a foundational advancement in the automation of data interpretation.

METHODS

MCBL model

In a dataset with N signals that were measured on a variety of samples, each signal S_i is modeled as the sum of the signal P_i from the sample, which typically involves a series of peaks, and the total background signal B_i. For each data point j in measurement i,

$$S_{i,j} = P_{i,j} + B_{i,j}. \quad (1)$$

Since in general B_i is composed of a unique mixture of K background signals, the background patterns and sample-specific weights are determined using a matrix factorization (MF) approach. The MF construction of the background model involves the matrix V, containing the collection of K signals from the background sources, and the matrix U, containing the amount of each background signal in each measured signal. The matrix product UV is thus the collection of total background signals for each measurement: $B \approx UV$. To create a model that does not require measurement of each background signal, which is typically not possible, the matrices U and V are learned from the measured spectra by considering the MF problem

$$S \approx UV. \quad (2)$$

In traditional implementations of matrix factorization, the residuals of the model, $R = S - UV$, are minimized with respect to ℓ_2 or similar loss metric. However, given that UV is the background model, R contains P, the signals of interest. In spectroscopic data, P is positive and can be large. Critically, large deviations are penalized heavily by traditional loss functions like ℓ_2. Therefore, this problem requires a novel approach to solving the matrix factorization problem, which allows for large deviations from the background model UV where signals of interest are present. If the signal of interest includes a peak (non-background signal) at the jth data

point in measurement i, then $R_{i,j}$ will be large and positive. While $R_{i,j}$ should be near zero for data points containing only background signal, the measurements S and thus the residuals R contain measurement noise. As a result, the distribution of $R_{i,j}$ values will be different when the signal of interest is absent or present. When absent, the measurement noise is typically well modeled by a Gaussian distribution, $\mathcal{N}_{\mu,\sigma}$. When present, the large residual intensities (peaks) are modeled by an exponential distribution, which when combined with the Gaussian distribution for noise yields the exponentially modified Gaussian (EMG) distribution:

$$\mathrm{EMG}_{\mu,\sigma,\lambda}(R_{ij}) = \frac{\lambda}{2} e^{\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2R_{ij})} \operatorname{erfc}\left(\frac{\mu+\lambda\sigma^2-R_{ij}}{\sqrt{2}\sigma}\right), \quad (3)$$

where erfc is the complementary error function, λ is the rate parameter of the exponential random variable, and μ and σ are the location and scale parameters of the Gaussian random variable, respectively. This distribution was previously used in biology,[21] psychology,[22] and finance.[23] Furthermore, the values λ, μ, and σ can vary along the measurement axis to increase the flexibility of the model, if required. In the present work we consider only a single σ and λ for a given dataset and fix the mean μ of the Gaussian noise to zero.

Since $\mathcal{N}_{\mu,\sigma}$ and the EMG distribution of Eq. (3) describe the distribution of residual intensities when the signal of interest is absent and present, respectively, a general expression for the distribution of residual intensities is their mixture:

$$\mathrm{EMGM}_{\mu,\sigma,\lambda,Z_{ij}}(R_{ij}) := (1-Z_{ij})\mathcal{N}_{\mu,\sigma}(R_{ij}) + Z_{ij}\,\mathrm{EMG}_{\mu,\sigma,\lambda}(R_{ij}), \quad (4)$$

where Z_{ij} indicates whether signal of interest is absent ($Z_{ij}=0$) or present ($Z_{ij}=1$) in the residual R_{ij}. Optimization of the matrix factorization model Eq. (2) corresponds to finding the background patterns, weights and distribution parameters such that the likelihood, corresponding to the product of Eq. (4) for all data points, is maximized. This enables U, V, Z, λ, μ, and σ to be learned concomitantly. A standard procedure in machine learning is to regularize optimization problems to make them well posed. In particular, to encourage the algorithm to find solutions with a small noise variance, we added a half normal prior on σ to regularize the optimization with respect to the parameter. The prior distribution has a variance σ_0^2 which can be used to control the strength of the regularization. This is necessary for the XRD dataset, since it does not include any substrate measurements. Therefore, we used $\sigma_0^2 = 0.01$ for the XRD dataset. Because the exact optimization of the binary variables $Z_{i,j}$ is computationally intractable, we employ an expectation-maximization algorithm.[24,25] Instead of inferring $Z_{i,j}$ directly, the algorithm computes the expected value $\mathbb{E}(Z_{ij})$, which is a continuous variable in the interval [0, 1]. From the equality

$$\mathbb{E}[Z_{ij}] = \mathbb{P}(Z_{ij}=1), \quad (5)$$

we also obtain the probability that the measured data point $S_{i,j}$ contains non-background signal. The algorithm for solving this implementation of probabilistic matrix factorization is described in ref. [26].

This approach to background identification enables unsupervised learning of the background model after choosing the value of a single parameter, the rank K of V, which corresponds to the number of background sources. While unsupervised methods for determining an appropriate value of K can be deployed,[27] we instead further constrain the matrix factorization model such that the results are relatively insensitive to K. This enables users to choose an upper bound for K and retain unsupervised operation. The constraints to the matrix factorization also enable semi-supervised operation, which enables both injection of prior knowledge of the background sources and deployment in data-starved situations where there are not enough examples of the background signals for the unsupervised model to robustly learn them. The constraints are implemented by defining kernel functions for each component of V. The most commonly used kernel is the squared exponential (SE) kernel, which enforces smoothness of each background signal. For example, the background model for the Raman data was obtained by using the SE kernel for all background components. Further, if the underlying physics of a given background signal give rise to a functional form or another physics-based constraint, this too can be used to constrain components in V. In fact, for the XRD dataset, the SE kernel was only used for two background signals. The other two background signals were constrained based on prior knowledge of the background signal from the crystalline substrate; the intensity of these background signals was constrained to zero except for the regions indicated above in the X-ray diffraction section. This is done with a simple projection: All values outside of the allowed ranges are set to zero in every gradient step of the optimization algorithm. Despite there only being one crystalline substrate, we used two vectors to express its signature to accommodate for any variations in this background signal over the set of measurements.

Library synthesis

The pseudo-ternary metal oxide composition gradient was fabricated using reactive direct current magnetron co-sputtering of Cu, Ca, and V metal targets in a non-confocal geometry onto a 100 mm diameter × 2.2 mm thick soda lime glass substrate with FTO coating (Tec15, Hartford Glass Company) in a sputter deposition system (Kurt J. Lesker, PVD75) at 10^{-5} Pa base pressure. The partial pressures of the deposition atmosphere containing inert sputtering gas Ar and reactive gas O_2 were 0.072 Pa and 0.008 Pa respectively. Deposition proceeded without active substrate heating, with the source powers set to 150 W, 11 W, and 95 W for the V, Cu, and Ca sources respectively. Deposition time per source was varied in order to achieve a total film thickness of 200 nm. The as-deposited composition library was annealed in a Thermo Scientific box oven in flowing air, with a 2 h ramp and 3 h soak at 550 °C, followed by passive cooling.

The 2121 samples forming the 15 pseudo-quaternary space composition library were deposited via inkjet printing onto 100 × 150 × 1.0 mm fluorine-doped tin oxide (FTO) coated boro-aluminosilicate glass (Corning Eagle XG Glass). The array of samples containing Mn, Fe, Ni, Cu, Co, and Zn was synthesized as a discrete library with 10 atom% composition steps in each element, using a print resolution of 2880 × 1440 dpi, as described previously.[28] Elemental precursor inks were prepared by mixing 3.33 mmoles of each metal precursor with 20 mL of stock solution. The stock solution of 500 mL 200 proof ethanol (Koptec), 16 mL glacial acetic acid (T.J. Baker, Inc.), 8 mL concentrated HNO_3 (EMD), and 13 g Pluronic F127 (Aldrich) was prepared beforehand. The metal precursors $Mn(NO_3)_2$ $4{\cdot}H_2O$ (0.88 g, 99.8%, Alfa Aesar), $Fe(NO_3)_3$ $9{\cdot}H_2O$ (1.43 g, 99.95%, Sigma Aldrich), $Co(NO_3)_2$ $6{\cdot}H_2O$ (0.93 g, 98%, Sigma Aldrich), $Ni(NO_3)_2$ $6{\cdot}H_2O$ (1.09 g, 98.5%, Sigma Aldrich), $Cu(NO_3)_2$ $3{\cdot}H_2O$ (0.83 g 99–104%, Sigma Aldrich), and $Zn(NO_3)_2$ $6{\cdot}H_2O$ (1.00 g 98%, Sigma Aldrich) were used as-received from the distributor. After inkjet printing, the inks were dried and converted to metal oxides by calcination in 0.395 atm O_2 at 450 °C for 10 h, followed by 0.395 atm O_2 at 750 °C for 10 h.

X-ray diffraction

XRD was performed on the pseudo-ternary metal oxide composition gradient using a Bruker DISCOVER D8 diffractometer, with a Bruker IμS source emitting Cu Kα radiation. Using a 0.5 mm collimator, the measurement area was approximately 0.5 mm × 1 mm. Within this measurement area the composition is uniform to with about 1 at.%. Measurements were taken on an array of 186 evenly spaced positions across the continuous composition library. Two-dimensional diffraction images taken by the VÅNTEC-500 detector were integrated into one-dimensional patterns using DIFFRAC.SUITE™ EVA software.

Raman spectroscopy

The 15 pseudo-quaternary composition space metal oxide sample library, was characterized using a Renishaw inVia Reflex Micro Raman spectrometer with Wire 4.1 software as described previously.[19] The instrument's laser wavelength was 532 nm, and the diffraction grating resolution 2400 lines mm^{-1} (visible). Spectra were taken over the range 67–1339.9 cm^{-1} using a ×20 objective. The Renishaw Streamline™ mapping system was used to automate spectral image collection in which a cylindrical lens-expanded 26 × 2 μm laser line was rastered over the measurement area. Spectra were acquired at 65 μm spatial resolution and 0.75 s exposure time.

DATA AVAILABILITY

The datasets analyzed during the current study are available in the Caltech Data repository: XRD at https://doi.org/10.22002/D1.1178, https://data.caltech.edu/records/1178 and Raman at https://doi.org/10.22002/D1.1179, https://data.caltech.edu/records/1179.

CODE AVAILABILITY

The codes pertaining to the current study will be available at http://www.cs.cornell.edu/gomes/udiscoverit/.

ACKNOWLEDGEMENTS

The development of the MCBL algorithm, inkjet printing synthesis, and Raman measurements were supported by a an Accelerated Materials Design and Discovery grant from the Toyota Research Institute. Initial design of the algorithm and data procurement were supported by the NSF Expedition award for Computational Sustainability CCF-1522054 and by Army Research Office (ARO) award W911-NF-14-1-0498. The implementation of the algorithm for automated, unsupervised operation was supported by MURI/AFOSR grant FA9550. Compute infrastructure was provided by NSF award CNS-0832782 and by ARO DURIP award W911NF-17-1-0187. The sputter deposition and XRD measurements were supported through the Office of Science of the U.S. Department of Energy under Award No. DE-SC0004993. The authors thank Edwin Soedarmadji for assistance with data management.

AUTHOR CONTRIBUTIONS

C.G. and J.G. identified the problem to be solved. S.A. and C.G. conceptualized the model. S.A. developed the mathematical framework, designed the algorithm, and implemented it. J.G., H.S. and D.G. inspected results. S.A., D.G. and J.G. created visualizations of the results. L.Z. performed materials synthesis and data acquisition for XRD data. J.H. synthesized materials for Raman measurements. D.B. and M.U. acquired and provided the Raman data. S.A., J.G., C.G., H.S. and D.G. wrote the paper.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Alberi, K. et al. The 2019 materials by design roadmap. *J. Phys. D* **52**, 013001 (2019).
2. Aspuru-Guzik, P. K. A. Alán. Report of the Clean Energy Materials Innovation Challenge Expert Workshop January 2018, Mission Innovation http://mission-innovation.net/wp-content/uploads/2018/01/Mission-Innovation-IC6-Report-Materials-Acceleration-Platform-Jan-2018.pdf.
3. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5 (2018).
4. Laue, M. Über die Interferenzerscheinungen an planparallelen Platten. *Ann. der Phys.* **318**, 163–181 (1904).
5. Seah, M. P. The quantitative analysis of surfaces by xps: a review. *Surf. Interface Anal.* **2**, 222–239 (1980).
6. Sonneveld, E. J. & Visser, J. W. Automatic collection of powder data from photographs. *J Appl. Crystallograph.* **8**, 1–7 (1975).
7. Tougaard, S. Algorithm for automatic X-ray photoelectron spectroscopy data processing and x-ray photoelectron spectroscopy imaging. *J. Vac. Sci. Technol.* **23**, 741–745 (2005).
8. Hattrick-Simpers, J. R., Gregoire, J. M. & Kusne, A. G. Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge. *APL Mater.* **4**, 053211 (2016).
9. Stein, H. S., Jiao, S. & Ludwig, A. Expediting combinatorial data set analysis by combining human and algorithmic analysis. *ACS Comb. Sci.* **19**, 1–8 (2017).
10. Tessier, F. & Kawrakow, I. Calculation of the electron–electron bremsstrahlung cross-section in the field of atomic electrons. *Nucl. Instr. Meth. Phys. Res. B* **266**, 625–634 (2008).
11. Kramers, H. A. Xciii. on the theory of x-ray absorption and of the continuous x-ray spectrum. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **46**, 836–871 (1923).
12. Davies, H., Bethe, H. A. & Maximon, L. C. Theory of Bremsstrahlung and pair production. II. Integral cross section for pair production. *Phys. Rev.* **93**, 788–795 (1954).
13. Bethe, H. A. & Maximon, L. C. Theory of Bremsstrahlung and pair production. I. Differential cross section. *Phys. Rev.* **93**, 768–784 (1954).
14. Tougaard, S. & Jorgensen, B. Inelastic background intensities in XPS spectra. *Surface Sci.* **143**, 482–494 (1984).
15. Zhao, J., Lui, H., McLean, D. I. & Zeng, H. Automated autofluorescence background subtraction algorithm for biomedical raman spectroscopy. *Appl. Spectrosc.* **61**, 1225–1232 (2007).
16. Markus, G., Konstantinos, N., Frank, P., Christian, M. & Andreas, O. Multivariate characterization of a continuous soot monitoring system based on Raman spectroscopy. *Aerosal Sci. Technol.* **49**, 997–1008 (2015).
17. Li, Z., Ludwig, A., Savan, A., Springer, H. & Raabe, D. Combinatorial metallurgical synthesis and processing of high-entropy alloys. *J. Mater. Res.* **33**, 3156–3169 (2018).
18. Zhao, J. Combinatorial approaches as effective tools in the study of phase diagrams and composition–structure–property relationships. *Prog. Mater. Sci.* **51**, 557–631 (2006).
19. Newhouse, P. F. et al. Solar fuel photoanodes prepared by inkjet printing of copper vanadates. *J. Mater. Chem. A* **4**, 7483–7494 (2016).
20. Wand, M. & Jones, M. Kernel Smoothing. New York: Chapman and Hall/CRC (1995).
21. Golubev, A. Exponentially modified gaussian (emg) relevance to distributions related to cell proliferation and differentiation. *J. Theor. Biol.* **262**, 257–266 (2010).
22. Palmer, E. M., Horowitz, T. S., Torralba, A. & Wolfe, J. M. What are the shapes of response time distributions in visual search? *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 58–71 (2011).
23. Carr, P., Madan, D. & Smith, H. R. Saddle point methods for option pricing. *J. Comput. Financ.* **13**, 49–61 (2009).
24. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–38 (1977).
25. Neal, R. M. & Hinton, G. E. *Learning in Graphical Models*. (MIT Press, Cambridge, 1999).
26. Ament, S., Gregoire, J. & Gomes, C. Exponentially-modified Gaussian mixture model: applications in spectroscopy. Preprint at arXiv:1902.05601 (2019).
27. Neal, R. M. Markov chain sampling methods for dirichlet process mixture models. *J. Comput. Graph. Stat.* **9**, 249–265 (2000).
28. Haber, J. A. et al. Discovering ce-rich oxygen evolution catalysts, from high throughput screening to water electrolysis. *Energy Environ. Sci.* **7**, 682–688 (2014).